MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Implementation and Evaluation of a Negation Tagger in a Pipeline-based System for Information Extraction from Pathology Reports

Kevin J. Mitchell [a], Michael J. Becich[a,b], Jules J. Berman[c], Wendy W. Chapman[b,d], John Gilbertson[a], Dilip Gupta[a], James Harrison[a,b,d], Elizabeth Legowski[a], Rebecca S. Crowley[a,b,d]

[a] *Centers for Pathology and Oncology Informatics, University of Pittsburgh*
[b] *Center for Biomedical Informatics, University of Pittsburgh*
[c] *National Cancer Institute, Bethesda MD*
[d] *Intelligent Systems Program, University of Pittsburgh*

Abstract

We have developed a pipeline-based system for automated annotation of Surgical Pathology Reports with UMLS terms that builds on GATE – an open-source architecture for language engineering. The system includes a module for detecting and annotating negated concepts, which implements the NegEx algorithm – an algorithm originally described for use in discharge summaries and radiology reports. We describe the implementation of the system, and early evaluation of the Negation Tagger. Our results are encouraging. In the key Final Diagnosis section, with almost no modification of the algorithm or phrase lists, the system performs with precision of 0.84 and recall of 0.80 against a gold-standard corpus of negation annotations, created by modified Delphi technique by a panel of pathologists. Further work will focus on refining the Negation Tagger and UMLS Tagger and adding additional processing resources for annotating free-text pathology reports.

Keywords:

Natural Language Processing, Information Extraction

Introduction

The Shared Pathology Informatics Network (SPIN) (http://spin.nci.nih.gov/) is a National Cancer Institute (NCI) sponsored cooperative agreement among four institutions (Harvard University, University of Indiana, UCLA, and University of Pittsburgh) to develop a model web-based system for accessing pathology data on archived human tissue specimens, across multiple institutions and databases. An important and difficult aspect of this work is the extraction of information (such as the diagnosis, findings, and relationship of tissue blocks to the specimen) from the free-text of Surgical Pathology reports. Information Extraction from pathology reports is complex. For example: (1) reports contain multiple sections (such as Final Diagnosis, Gross Description, Comment, etc) that vary in narrative structure and uniformity (2) there is institutional variation in reporting practices (such as differences in the keywords that delimit important sections of the report), and (3) reports contain negative as well as positive findings and diagnoses. We describe the early development of a pipeline-based system for machine annotation of surgical pathology reports. Example phrases from surgical pathology reports are shown in Table 1. The system has been used to automatically annotate 20,000 randomly selected, surgical pathology reports from our institution. De-identified reports and annotations are available for retrieval by all SPIN institutions through a peer-to-peer network available developed for SPIN [1]. This report focuses on the early evaluation of a particularly critical component of the system – the Negation Tagger. The modular nature of the pipeline approach simplifies the evaluation of discrete processes so that they can be tested and refined independently of the remaining components.

Materials and Methods

Materials. The system was developed using GATE [2,3] – a software architecture for language engineering which includes an architecture describing the relationship of language processing components, a framework of Java classes for many of these components, and a development environment for creating language engineering (LE) applications. GATE is available from the University of Sheffield (http://gate.ac.uk/) under the terms of the GNU General Public License. GATE is designed to provide modularity and flexibility by delineating components as (1) Language Resources – documents and corpora of documents, (2) Processing Resources – discrete processing components or sequences of components arranged in pipelines, or (3) Visual Resources – graphical environments for development and testing of LE applications that also provide interfaces for human annotation in order to create reference or gold standard document sets. GATE provides:

- *Off-the-shelf processing resources that need little or no modification*: (e.g. Annotation Reset - clears existing annotations, and Tokeniser - parses words, numbers, and punctuation).
- *configurable GATE processing resources* (e.g. Gazetteer lists to annotate documents based on keyword lists).
- *a GATE integrated pattern engine* (e.g. JAPE - a multi-pass regular expression parser tightly integrated with

Table 1: Examples phrases and text from corpus by response category

Response Category	*Example sentences or phrases with gold-standard negation enclosed in < > and automated negation **bolded***
True Positive	"BILATERAL SEMINAL VESICLES ARE FREE OF <**CARCINOMA**>."
	"NO EVIDENCE OF <**DYSPLASIA**> OR <**MALIGNANCY**>."
	"NEGATIVE FOR <**INTESTINAL METAPLASIA**>, <**DYSPLASIA**> AND <**HELICOBACTER PYLORI**>."
Partially Positive	"No <**vascular** invasion> is seen."
	"NO SIGNIFICANT<**PATHOLOGIC** CHANGE>."
	"There is no <specialized **intestinal metaplasia**>
False Positive	"No stones are **present**."
	"LESION, SINONASAL TRACT, SIDE NOT INDICATED: -- **ADENOID CYSTIC CARCINOMA**"
	"The lack of HHV-8 does not rule out the possibility of **Castleman' s disease**."
False Negative	"NO <PERINEURAL INVASION>."
	"There is however no evidence of cryptitis, crypt abscesses, significant intraepithelial inflammation, nor a <thickened collagen layer>."
	"<HELICOBACTER PYLORI> ORGANISMS NOT SEEN."

GATE's document annotation representation). Developers use JAPE transducers to configure sets of parsing rules that work with arbitrary annotation sets.

System Architecture. Using GATE resources we configured a Corpus Pipeline to: (1) annotate the sections of the surgical pathology report (e.g. Final Diagnosis, Gross Description, Comment), (2) annotate concepts using a subset of UMLS semantic types, and (3) differentially annotate negated concepts. The entire processing sequence is shown in Figure 1. Two aspects of this sequence require further clarification:

UMLS Concept Annotation. Moving from left to right in a report line, words are grouped into phrases incrementing in length from one to four. Each subset is then matched against the UMLS, and matched phrases are annotated as concepts with Concept Unique Identifiers (CUIs), concept name, and semantic type included as features of the annotation. Look-ups are performed via Java Remote Method Invocation (RMI) to the NLM Knowledge Source Server (http://umlsks1.nlm.nih.gov/kss), using an exact match criteria on all sub-strings. Later in the pipeline, the Filter Processing Resource removes concepts that are completely subsumed by longer concepts, as well as common stop words (e.g. and, of, the). Stop words are annotated earlier in the pipeline during Gazetteer look-up (Figure 1). In the current version, decomposition of look-up and filter, sacrifices processing time in favor of modularity and rapid application development. To limit spurious (false positive) tagging we restricted tagging to 35 semantic types chosen from 5 semantic categories relevant to surgical pathology reports. A list of all UMLS semantic types used is available at http://spin.nci.nih.gov/content.

Negated Concept Annotation Our Negation Tagger implements NegEx [4-6] a regular-expression based algorithm for detecting negation, with a reported precision of 0.85. UMLS Concepts are tagged as negated if they fall six elements after a pre-negation phrase or six elements before a post-negation

phrase. Elements may be either (a) single words or (b) phrases that match to UMLS concepts. Importantly, the algorithm excludes pseudo-negations - phrases containing negations that must be ignored (i.e. "impossible to rule out"). Pre-negation phrases, post-negation phrases, and pseudo-negation phrases are annotated earlier in the pipeline during Gazetteer look-up (Fig 1).

Development of a Gold-Standard Corpus using a modified Delphi Technique. In order to specifically evaluate the performance of the Negation Tagger, four pathologists used the manual annotation interface in GATE to mark-up pathology reports, indicating all negated concepts within the document set.

Measurement Study: We first performed a measurement study using 130 randomly selected reports and four pathologists, in order to determine the inter-rater reliability for coding of negated concepts. In this study, annotators worked through de-identified reports, in a single pass, identifying all phrases that they considered to be Negated Concepts. We computed the inter-rater reliability and estimated the reliability as a function of number of raters using the Spearman-Brown prophecy formula [7] (an estimate for total number of concepts was derived from a single annotator). On the basis of the results (see results section), we elected to perform our demonstration study using a Modified Delphi technique to achieve consensus among the panel, rather than by using a single pass panel.

Demonstration Study: We asked four pathologists to annotate 250 randomly selected surgical pathology reports from a single multi-hospital medical center. Annotators were trained over several group meetings by working through multiple examples together using a set of written criteria for identifying negation.

In this study, annotators worked through de-identified reports in two passes. In the first pass, all document text was read and negated concepts were annotated. Between the first and second pass, a merged file was created that contained only those annotations in which there was disagreement among the judges. In the second pass, annotators re-coded these disagreements. Following the second pass, a merged file was created that contained remaining disagreements. The remaining disagreements were resolved by a subset of the annotators after group discussion. We computed inter-rater reliability and determined frequencies for changing annotations from first to second pass. The final gold-standard corpus contained 250 documents with a total of 11449 non-blank lines, 65858 words, and a total of 311 human-panel-coded negations.

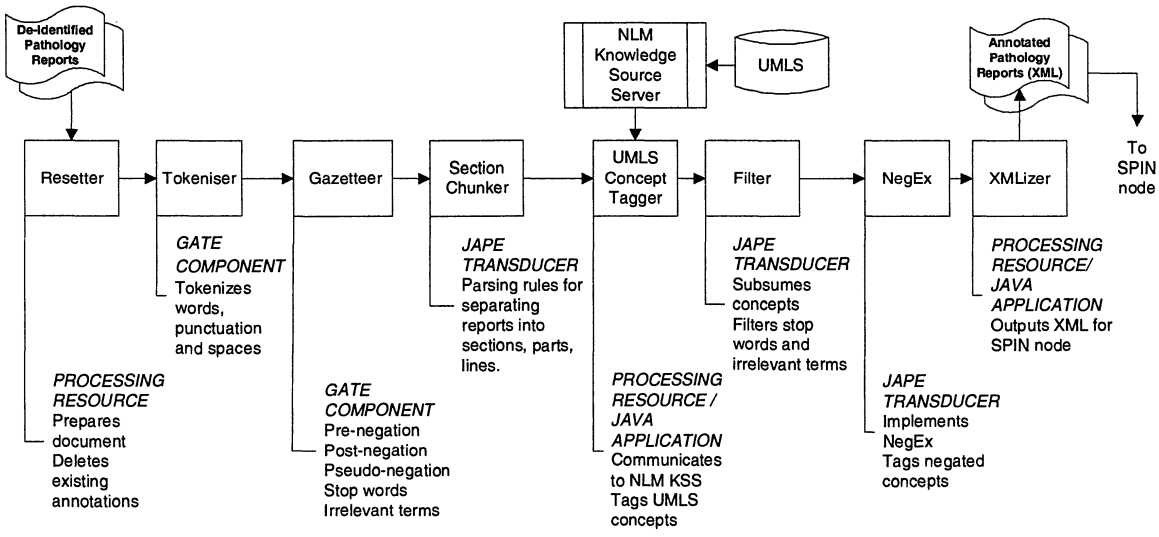

Figure 1 - Pipeline System for Annotation of Surgical Pathology Reports

Comparison of Automated Annotations to Gold-Standard. We computed precision and recall of our automated Negation Tagger, compared to the gold standard, using precise, lenient and average metrics. Precise metrics only consider annotations equivalent when they completely overlap (co-extensive). Lenient metrics allow any overlap. Definitions of precision and recall under these conditions are shown below – where PP denotes partial overlap, from key to response document [8].

Strict Precision (SP) = $TP/(TP+FP+\frac{1}{2}PP)$
Strict Recall (SR) = $TP/(TP+FN+\frac{1}{2}PP)$
Lenient Precision (LP) = $(TP+\frac{1}{2}PP)/(TP+FP+\frac{1}{2}PP)$
Lenient Recall (LR) = $(TP+\frac{1}{2}PP)/(TP+FN+\frac{1}{2}PP)$
Average Precision = $(SP + LP)/2$
Average Precision = $(SR + LR)/2$

We computed metrics over all report sections and separately for each report section to identify performance differences among sections. We also examined the false negatives - phrases that were marked as negated in the gold-standard but not by our system, to determine whether the failure related to: (1) a mismatch between concepts tagged by humans and those tagged by our UMLS tagger or (2) failure of the negation tagger to identify a concept as negated. It is important to know the relative distribution of these false negative cases – only cases in the second category reflect a failure or the Negation Tagger. To determine the percentage of false positives that were attributable to the NegEx algorithm as opposed to the UMLS tagger, we used the human annotated concepts from the gold-standard to augment the automated UMLS tagger, and the resulting document set was then run through the standard Negation Tagger.

Result

Measurement Study. The overall reliability among the four pathologists was 67% (mean % agreement). The estimated relationship between number of annotators and reliability suggest that even by doubling the number of judges, the reliability increases only to 80%. On further investigation, approximately half of the discrepancies were related to 'misses' – cases where a straightforward instance of negation in the text was simply not recognized by the annotator. Based on the findings of our measurement study, we elected to create a gold-standard corpus of negation annotations using a modified Delphi technique, in which annotators have at least one additional chance to catch misses and correct over-calls.

Demonstration Study. In round one, mean % agreement across all annotators was 69.9% (nearly identical to the measurement study reliability). All four annotators agreed on 36.5% of the negations (N=140). But 63.5% of negations (N=244) were discrepant among one or more annotators. Only these negations were presented in round two. Following the second round, there was 86.1% mean agreement. All four annotators now agreed on 73.4% of negations (N=282) and 26.6% of negations (n=102) were discrepant. The majority of these remaining discrepancies did not fit the criteria that were established by the group for negation and were removed during final review (N=73). Table 2 shows the changes by annotator as a result of the second pass. Importantly, we noted significant variation was seen for rates of missing and over-marking of negations, affirming the need for multiple annotators. The final gold-standard corpus included 250 reports with 311 human-coded negation annotations.

Comparison of Automated Annotations to Gold-Standard. We then determined the precision and recall of negation annotations generated by our system, against negation annotations in the gold-standard corpus. Table 3 shows the total number of gold-standard and automated annotations, true-positives (correct), partial-positives (partially correct), false- positives (spurious), true-negatives (missing), precision and recall, under lenient, strict and average conditions, by report section and overall. Performance is best within the key Final Diagnosis section

Table 2: Outcome of second pass annotation for discrepant negations from 1st round (N=244)

Annotator	Negated both rounds	Only negated in 1st round (overmarked)	Only negated in 2nd round (miss)	Not negated either round	Changed from round 1 to 2
1	61.9% (n=151)	2.0% (n=5)	7.0% (n=17)	29.1% (n=71)	9.0% (n=22)
2	53.7% (n=131)	2.9% (n=7)	11.9% (n=29)	31.6% (n=77)	14.8% (n=36)
3	31.6% (n=77)	11.5% (n=28)	38.9% (n=95)	18.0% (n=44)	50.4% (n=123)
4	19.3% (n=47)	16.4% (n=40)	29.5% (n=72)	34.8% (n=85)	45.9% (n=112)
Average	**41.6%**	**8.2%**	**21.8%**	**28.4%**	**30.0%**

Table 3: Performance Metrics across report sections for automated system against Gold-Standard

Report Section	Negations (N)		Tallies				Precision			Recall		
	Gold	*Auto*	TP	PP	FP	FN	*Lenient*	*Average*	*Strict*	*Lenient*	*Average*	*Strict*
Gross Description	55	42	15	23	4	20	0.87	0.68	0.49	0.57	0.45	0.32
Final Diagnosis	149	144	106	22	16	23	0.88	0.84	0.80	0.84	0.80	0.76
Microscopic Description	17	35	3	11	21	7	0.29	0.19	0.10	0.55	0.37	0.19
Comment	45	34	10	14	10	23	0.63	0.50	0.37	0.42	0.34	0.25
Other	45	28	1	3	34	31	0.068	0.047	0.027	0.074	0.052	0.029
All sections	311	283	138	74	71	110	0.71	0.64	0.56	0.61	0.55	0.48

of the report (Precision = 0.84, Recall = 0.8), and worst within the Microscopic Description section (Precision = 0.19, Recall = 0.37). Final Diagnosis fields are present in all Surgical Pathology reports, and are typically highly structured and formalized text, often in outline form. In contrast, the Microscopic description section is typically more complex narrative, and often only used in a subset of cases to explain diagnostic reasoning in difficult cases. After augmenting the UMLS Tagger with human-coded concepts, we measured the precision and recall and determined the remaining false negatives for each reports section and overall. In this condition we are testing the behavior of the Negation algorithm separately from the UMLS Tagger. Average precision rises from 0.64 to 0.77 and average recall rises from 0.55 to 0.83. In the Final Diagnosis section, average recall and precision reach 0.90 and 0.98 respectively. As shown in Table 4, the majority of false negatives can be attributed to the tagging of concepts by the UMLS Tagger.

Discussion

We have described the early development of a pipeline-based system for annotating surgical pathology reports. The system implements a simple method for detecting and annotating UMLS concepts as well as annotating negations based on the NegEx algorithm. We tested the ability of our system to accurately annotate negations against a gold-standard corpus of negation annotations created by four pathologists. Overall precision is lower than that reported by Chapman et al[4], using the NegEx algorithm. However, NegEx was not specifically developed for this class of medical documents. Consequently the negation and pseudo-negation phrases that are used may not adequately cover the spectrum of phrases used by pathologists in their reports.

On the other hand, our results demonstrate that NegEx performs reasonably well within the Final Diagnosis section, as opposed to other sections, even without significant changes to the technique. Better performance in the Final Diagnosis section most likely reflect the simple linguistic constructions common in this report section. Fortunately, the Final Diagnosis section contains most of the major diagnoses and findings that we are seeking to extract from report text for the SPIN project. Further refinements to our Negation Tagger will focus on expanding the set of negation and pseudo-negation phrases, specific to those commonly used in surgical pathology reports.

Report Section	Total FN	FN related to UMLS Tagger	FN related to Negation Tagger
Gross Description	23	18	5
Final Diagnosis	25	21	4
Microscopic Description	7	4	3
Comment	23	4	19
Other	31	9	22
All Sections	110	58	52

Table 4. Breakdown of false negatives by report section

Our underlying architecture builds on a pipeline system developed in GATE – an open source system for developing Language Engineering appplications. GATE has significant advantages for creating and iteratively testing components in a complex information extraction system.

First, the modular approach enables components of the system to be tested semi-independently – resulting in iterative system improvement. In this study, we determined the performance of a critical component of the system - the Negation Tagger - within the context of the entire pipeline. By transferring a particular annotation set from the human coded gold-standard, we determined the percentage of false negatives attributable to the UMLS Tagger separate from the Negation Tagger. False Negatives related to the UMLS Tagger may be due to (1) a failure of the UMLS Tagger to correctly map the text phrase to the UMLS concept, (2) concepts which appear in report text but which are missing from the UMLS (e.g. "No <perineural invasion>") or from the restricted semantic types, or (3) concepts annotated by humans that only partially overlap with UMLS concepts, often

because humans encode longer, composite phrases including concept modifiers (e.g. "There is no <specialized intestinal metaplasia>").

Given the relative proportion of false negatives attributable to the UMLS Tagger, improvements to this part of the system may prove to be the most efficient way to improve the overall performance in annotation of negations. In future work, we plan to integrate more sophisticated methods of UMLS concept mapping (such as UMLS MetaMap) as well as parts-of-speech tagging into our pipeline.

A second benefit of the GATE architecture is that it enables human annotation of documents using the same annotation schema used for machine annotation. In this study, we used existing GATE resources to develop our annotated corpus, and for merging annotations from multiple annotators.

The software for the entire pipeline, including the Negation Tagger, is available under terms of the GNU General Public License at http://spin.nci.nih.gov/content along with instructions for implementing it within the GATE framework.

Acknowledgments

This work was supported by National Cancer Institute Grant UO1 CA91343 for the Shared Pathology Informatics Network (SPIN).

References

[1] Holzbach A, Cheuh H, Porter AJ, Kohane IS and Berkowicz D. A query engine for distributed medical databases. Accepted to *MedInfo 2004.*

[2] Cunningham H, Wilks Y, and Gaizauskas R. GATE -- a General Architecture for Text Engineering. In: *Proceedings of the 16th Conference on Computational Linguistics* (COLING-96), 1996.

[3] Cunningham H, Humphreys, K, Gaizauskas R, and Wilks Y. Software Infrastructure for Natural Language Processing. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing* (ANLP-97), 1997.

[4] http://omega.cbmi.upmc.edu/~chapman/NegEx.html

[5] Chapman W, Bridewell W, Hanbury P, Cooper G, and Buchanan B. Evaluation of negation phrases in narrative clinical reports. In *Proc AMIA Symp* 105–9, 2001.

[6] Chapman W, Bridewell W, Hanbury P, Cooper G, and Buchanan B. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–10, 2001.

[7] Spearman C. Correlation calculated with faulty data. *British Journal of Psychology*, 3, 271-295, 1910.

[8] Douthat A. The Message Understanding Conference Scoring Software User's Manual. In Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

Address for correspondence

Kevin J. Mitchell MS or Rebecca S. Crowley MD MS,
Centers for Pathology and Oncology Informatics,
UPMC Shadyside 5230 Centre Avenue,
Pittsburgh PA 15232.
mitchellkj@msx.upmc.edu
or crowleyrs@msx.upmc.edu